



# The UPM RT09 Meetings Evaluation System

**Jose M. Pardo, Roberto Barra, Beatriz  
Martínez**

Speech Technology Group. ETSI Telecomunicación.  
Universidad Politécnica Madrid.

# Contents

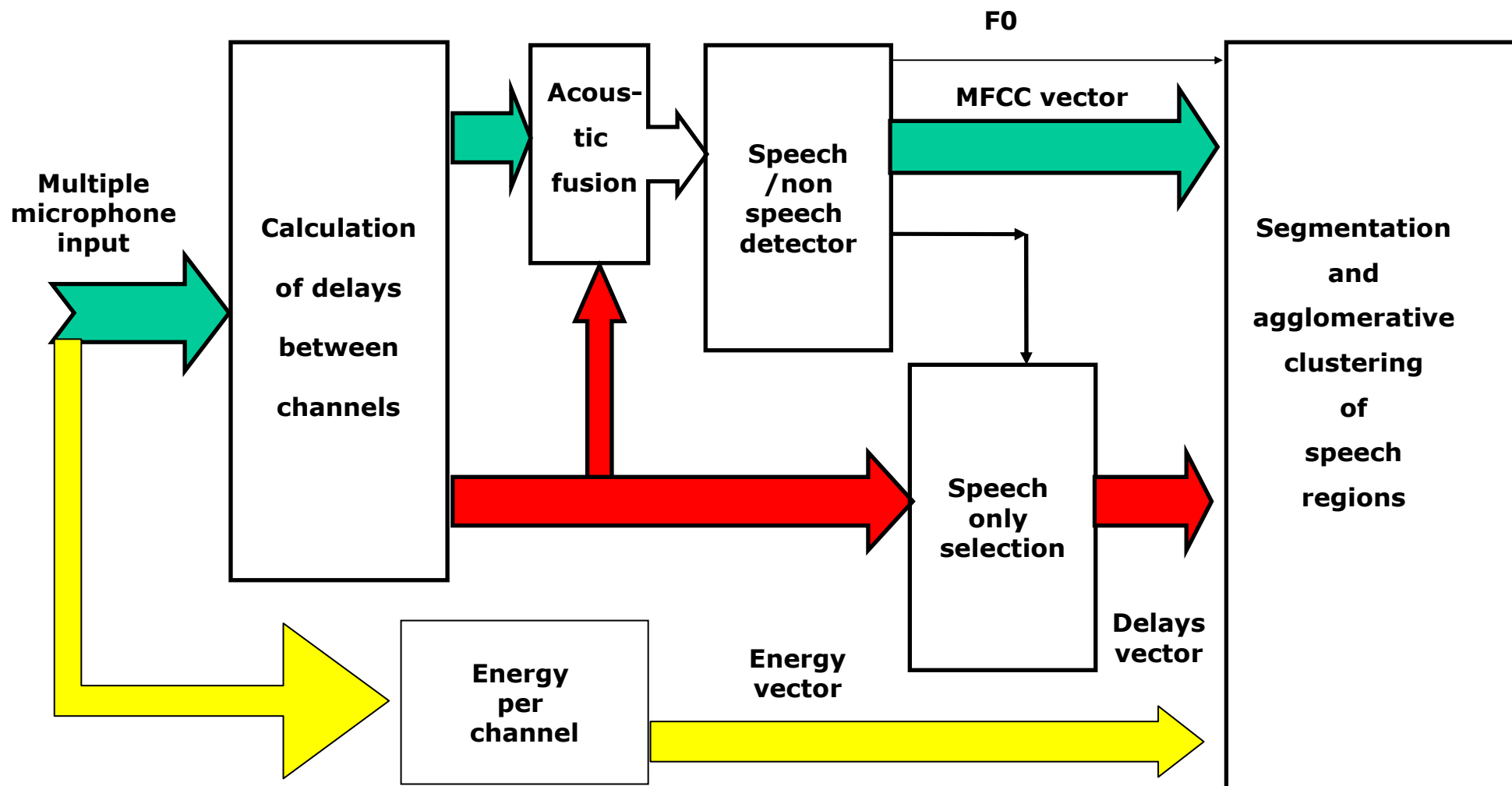
- **Introduction**
- **System description**
  - Channels preprocessing
  - Acoustic fusion and parameters extraction
  - Speech/non speech detection
  - Segmentation and clustering
  - Frame purification
  - Stopping criterion
  - Improvements using extra features
- **Results**
- **After evaluation analysis**
- **Conclusion**

# Introduction

- First time UPM participates
- Start from system submitted from ICSI in 2006 [1] named c-spsnspdelay adding frame purification
- Add extra features: normalized energy for every channel and F0

[1] X. Anguera, C. Wooters, J. M. Pardo, "Robust speaker diarization for meetings: ICSI RT06s meetings evaluation system," Lecture Notes in Computer Science, Volume 4299/2006, pp. 346-358, ISSN 0302-9743, 2006

# System description



# System description: Channels preprocessing

- Noise reduction (Wiener filtering)- Qualcomm-ICSI-OGI
- Time delay of arrival (TDOA) estimation, Generalized cross correlation with phase transform (GCCPHAT) plus peak selection algorithm. Creation of a TDOA vector every 10 msec.
- Estimating normalized energy for each channel at frame n

$$ene[n, i] = 10\log_{10}\left(\frac{1}{n_2 - n_1} \sum_{k=n_1}^{n_2} s_i^2[k]\right)$$

$$\overline{ene}[n, i] = \frac{ene[n, i]}{\sum_{i=0}^{D_{ene}-1} ene[n, i]}$$

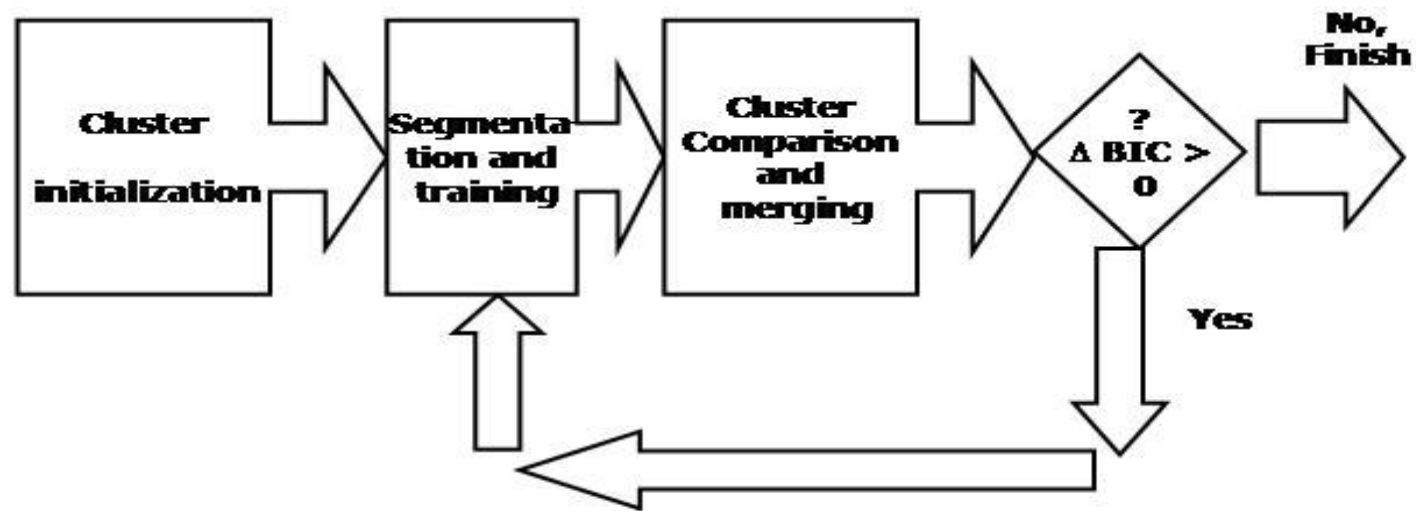
# Acoustic fusion and feature extraction

- Beamforming a single channel : delay and sum (Beamformit v 2.0 by X. Anguera), 500 msec window.
- MFCC extraction 19th order, 30 msec window, 10 msec shift.
- Calculate log fundamental frequency, F0. We interpolate it for unvoiced segments

# Speech- non speech detector

- Same as presented by ICSI RT06
- Initial segmentation based on energy and minimum number of samples for non-speech (1600- 0.15 sec) . Used to estimate initial non speech models
- Two GMM models, one for speech, one for non-speech, parameters: minimum duration (0.7 sec) and number of initial mixtures for speech (2).

# Segmentation and agglomerative clustering



# Segmentation and agglomerative clustering

- Initial gaussians per model (5). They are increased after merging clusters
- Minimum duration (2.5 sec)
- Decision on segmentation

$$\max_{\theta_i} \log p(x[n] | \theta_i)$$

- Initialization= homogenous segmentation, 16 initial clusters

# Frame purification

- Some percentage of frames (silences, noises) are too short to be part of a new cluster but corrupt the cluster models [2].
- Aims to detect and eliminate non-speech frames that do not help in discriminating speakers
- 10% of frames with highest likelihood computed on gaussians with smaller variance are removed for training models that have more than two gaussians before computing  $\Delta\text{BIC}$  (merging criterion)

[2] X. Anguera. "Robust speaker diarization for meetings", Ph D Thesis, Universitat Politècnica de Catalunya, October 2006

# Merging

- Delta BIC without penalty term of standard BIC[2]

$$\Delta BIC = \log p(D | \theta) -$$

$$\log p(D_a | \theta_a) - \log p(D_b | \theta_b)$$

*n parameter in  $\theta$  ~~n~~ parameter in  $\theta_a$  +*

*n parameter in  $\theta_b$*

[2] J. Ajmera, C. Wooters : A Robust speaker clustering algorithm, IEEE ASRU 2003.

# Using new features



MFCC features,  
starting with 5  
Gaussians



Concatenation  
of TDOA plus  
energy 1  
Gaussian



Log F0 1  
Gaussian



Training separate  
models for the  
same cluster

## Using extra features

- Compound likelihood as in [3]

$$\begin{aligned} \log p(x[n], y[n], z[n] | \theta_a) = \\ \alpha \log p(x[n] | \theta_{ax}) + \beta \log p(y[n] | \theta_{ay}) + \\ \gamma \log p(z[n] | \theta_{az}) \\ \alpha + \beta + \gamma = 1 \end{aligned}$$

- Used both in segmentation and in merging (for  $\Delta$  BIC)

[3] J.M. Pardo, X. Anguera, C. Wooters, "Speaker Diarization for Multiple-Distant-Microphone Meetings Using Several Sources of Information" IEEE Transactions on Computers, Vol. 56, No. 9, September 2007, pp 1212-1224

# Contrastive system

- Instead of three streams, use four streams of data with the same philosophy



MFCC features,  
starting with 5  
Gaussians



TDOA features  
starting with 1  
Gaussian



Energy features  
starting with 1  
Gaussian



Log F0 starting  
with 1 Gaussian

Training separate  
models for the  
same cluster

# Results

System	Official DER Results	Relative improvement	DER for RT06 plus devel06	Relative improvement
<i>p-gthtls</i>	21.38%	4.6%	10.81% ± 0.047	2.17%
<i>c-gthmdef</i>	22.43%		11.05% ± 0.047	

# After evaluation analysis

System	DER for RT09	Relative improvement from the baseline
<i>Base RT06 system</i>	<b>25.67 %</b>	
<i>Base plus energy features</i>	<b>23.64 %</b>	<b>7.9%</b>
<i>Base plus F0 feature</i>	<b>22.94 %</b>	<b>10.63%</b>
<i>All included (official result)</i>	<b>21.38 %</b>	<b>16.71%</b>

- Difference is significant with 95% confidence interval

## Development analysis: Comparison using energy features

	<b>All06 (rt06 plus devel 06 :20 meetings)</b>	<b>RT 07</b>	<b>RT 09</b>
<i>RT 06 system</i>	13.4	14.12	25.63
<i>Baseline plus energy</i>	12.7	13.61	23.64

# Comparison using F0 analysis and energy

	RT 06	RT09
<i>Base line</i>	18.52	25.67
<i>Base line plus F0</i>	15.63	22.94
<i>Base line plus enery plus F0</i>	14.86	21.38

# Conclusion

- We presented an improved method to do speaker diarization.
- We added a normalized energy channels vector
- We added a log F0 vector
- We obtained 16.71 % improvement over the baseline